

From narrative descriptions to MedDRA: automagically encoding adverse drug reactions

Carlo Combi^a, Margherita Zorzi^a, Gabriele Pozzani^b, Ugo Moretti^b

^a*Department of Computer Science, University of Verona, Italy*

^b*Department of Diagnostics And Public Health, University of Verona, Italy*

Abstract

Context The collection of narrative spontaneous reports is an irreplaceable source for the prompt detection of suspected adverse drug reactions (ADRs): qualified domain experts manually revise a huge amount of narrative descriptions and then encode texts according to **MedDRA** standard terminology. The manual annotation of narrative documents with medical terminology is a subtle and expensive task, since the number of reports is growing up day-by-day.

Objectives Natural Language applications can support the work of people responsible for pharmacovigilance. Our objective is to develop Natural Language Processing (NLP) algorithms and tools oriented to the healthcare domain, in particular to the detection of ADR clinical terminology. Efficient applications can concretely improve the quality of the experts' revisions: NLP software can quickly analyze narrative texts and offer a (as much as possible) correct solution (a list of **MedDRA** terms) that the expert has to revise and validate.

Methods **MagiCoder**, a Natural Language Processing algorithm, is proposed for the automatic encoding of free-text descriptions into **MedDRA** terms. **MagiCoder** procedure is efficient in terms of computational complexity (in particular, it is linear in the size of the narrative input and the terminology). We tested it on a large dataset of about 4500 manually revised reports, by performing an automated comparison between human and **MagiCoder** revisions.

Results For the current base version of **MagiCoder**, we measured: on short descriptions, an average recall of 86% and an average precision of 88%; on medium-long descriptions (up to 255 characters), an average recall of 64% and an average precision of 63%.

Conclusions From a practical point of view, **MagiCoder** reduces the time required for encoding ADR reports. Pharmacologists have simply to review and validate the **MedDRA** terms proposed by the application, instead of choosing the right terms among the 70K low level terms of **MedDRA**. Such improvement in the efficiency of pharmacologists' work has a relevant impact also on the quality of the subsequent data analysis. We developed **MagiCoder** for the Italian pharma-

Email addresses: carlo.combi@univr.it (Carlo Combi), margherita.zorzi@univr.it (Margherita Zorzi), gabriele.pozzani@univr.it (Gabriele Pozzani), ugo.moretti@univr.it (Ugo Moretti)

covigilance language. However, our proposal is based on a general approach, not depending on the considered language nor the term dictionary.

Keywords: Natural Language Processing, Healthcare informatics, Pharmacovigilance, Adverse Drug Reactions, Term identification

1. Introduction

Pharmacovigilance includes all activities aimed to systematically study risks and benefits related to the correct use of marketed drugs. The development of a new drug, which begins with the production and ends with the commercialization of a pharmaceutical product, considers both pre-clinical studies (usually tests on animals) and clinical studies (tests on patients). After these phases, a pharmaceutical company can require the authorization for the commercialization of the new drug. Notwithstanding, whereas at this stage drug benefits are well-known, results about drug safety are not conclusive [1]. The pre-marketing tests cited above have some limitations: they involve a small number of patients; they exclude relevant subgroups of population such as children and elders; the experimentation period is relatively short, less than two years; the experimentation does not deal with possibly concomitant pathologies, or with the concurrent use of other drugs. For all these reasons, non-common Adverse Drug Reactions (ADRs), such as slowly-developing pathologies (e.g., carcinogenesis) or pathologies related to specific groups of patients, are hardly discovered before the commercialization. It may happen that drugs are withdrawn from the market after the detection of unexpected collateral effects. Thus, it stands to reason that the post-marketing control of ADRs is a necessity, considering the mass production of drugs. As a consequence, pharmacovigilance plays a crucial role in human healthcare improvement [1].

Spontaneous reporting is the main method pharmacovigilance adopts in order to identify adverse drug reactions. Through spontaneous reporting, health care professionals, patients, and pharmaceutical companies can voluntarily send information about suspected ADRs to the national regulatory authority¹. The spontaneous reporting is an important activity. It provides pharmacologists and regulatory authorities with early alerts, by considering every drug on the market and every patient category.

The Italian system of pharmacovigilance requires that in each local healthcare structure (about 320 in Italy) there is a qualified person responsible for pharmacovigilance. Her/his assignment is to collect reports of suspected ADRs and to send them to the National Network of Pharmacovigilance (RNF, in Italian) within seven days since they have been received². Once reports have been notified and sent to RNF they are analysed by both local pharmacovigilance

¹in Italy, the Drug Italian Agency AIFA – Agenzia Italiana del FArmaco, <http://www.agenziafarmaco.gov.it/>

²According to the Italian Law, Art. 132 of Legislative Decree Number 219 of 04/24/2006.

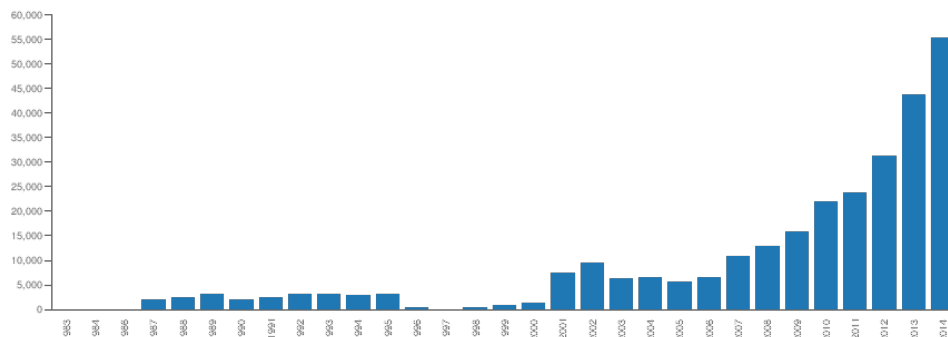


Figure 1: The yearly increasing number of reports about suspected adverse reactions induced by drugs in Italy.

centres and by the Drug Italian Agency (AIFA). Subsequently, they are sent to Eudravigilance [2] and to VigiBase [3] (the European and the worldwide pharmacovigilance network RNF is part of, respectively). In general, spontaneous ADR reports are filled out by health care professionals (e.g., medical specialists, general practitioners, nurses), but also by citizens. In last years, the number of ADR reports in Italy has grown rapidly, going from approximately ten thousand in 2006 to around sixty thousand in 2014 [4], as shown in Figure 1.

Since the post-marketing surveillance of drugs is of paramount importance, such an increase is certainly positive. At the same time, the manual review of the reports became difficult and often unbearable both by people responsible for pharmacovigilance and by regional centres. Indeed, each report must be checked, in order to control its quality; it is consequently encoded and transferred to RNF via “copy by hand” (actually, a printed copy).

Recently, to increase the efficiency in collecting and managing ADR reports, a web application, called **VigiFarmaco**³, has been designed and implemented for the Italian pharmacovigilance. Through **VigiFarmaco**, a spontaneous report can be filled out online by both healthcare professionals and citizens (through different user-friendly forms), as anonymous or registered users. The user is guided in compiling the report, since it has to be filled step-by-step (each phase corresponds to a different report section, i.e., “Patient”, “Adverse Drug Reaction”, “Drug Treatments”, and “Reporter”, respectively). At each step, data are validated and only when all of them have been correctly inserted the report can be successfully submitted.

Once ADR reports are submitted, they need to be validated by a pharmacovigilance supervisor. **VigiFarmaco** provides support also in this phase and is useful also for pharmacovigilance supervisors. Indeed, **VigiFarmaco** reports are high-quality documents, since they are automatically validated (the pres-

³Available at <https://www.vigifarmaco.it>

ence, the format, and the consistency of data are validated at the filling time). As a consequence, they are easier to review (especially with respect to printed reports). Moreover, thanks to **VigiFarmaco**, pharmacologists can send reports (actually, XML files [5]) to RNF by simply clicking a button, after reviewing it.

Online reports have grown up to become the 30% of the total number of Italian reports. As expected, it has been possible to observe that the average time between the dispatch of online reports and the insertion into RNF is sensibly shorter with respect to the insertion from printed reports. Notwithstanding, there is an operation which still requires the manual intervention of responsables for pharmacovigilance also for online report revisions: the encoding in **MedDRA** terminology of the free text, through which the reporter describes one or more adverse drug reactions. **MedDRA** (Medical Dictionary for Regulatory Activities) is a medical terminology introduced with the purpose to standardize and facilitate the sharing of information about medicinal products in particular with respect to regulatory activities [6]. The description of a suspected ADR through narrative text could seem redundant/useless. Indeed, one could reasonably imagine sound solutions based either on an autocompletion form or on a menu with **MedDRA** terms. In these solutions, the description of ADRs would be directly encoded by the reporter and no expert work for **MedDRA** terminology extraction would be required. However, such solutions are not completely suited for the pharmacovigilance domain and the narrative description of ADRs remains a desirable feature, for at least two reasons. First, the description of an ADR by means of one of the seventy thousand **MedDRA** terms is a complex task. In most cases, the reporter who points out the adverse reaction is not an expert in **MedDRA** terminology. This holds in particular for citizens, but it is still valid for several professionals. Thus, describing ADRs by means of natural language sentences is simpler. Second, the choice of the suitable term(s) from a given list or from an autocompletion field can influence the reporter and limit her/his expressiveness. As a consequence, the quality of the description would be also in this case undermined. Therefore, **VigiFarmaco** offers a free-text field for specifying the ADR with all the possible details, without any restriction about the content or strict limits to the length of the written text. Consequently, **MedDRA** encoding has then to be manually implemented by qualified people responsible for pharmacovigilance, before the transmission to RNF. As this work is expensive in terms of time and attention required, a problem about the accuracy of the encoding may occur given the continuous growing of the number of reports.

According to the described scenario, in this paper we propose **MagiCoder**, an *original* Natural Language Processing (NLP) [7] algorithm and related software tool, which automatically assigns one or more terms from a dictionary to a narrative text. A preliminary version of **MagiCoder** has been proposed in [8]. **MagiCoder** has been first developed for supporting pharmacovigilance supervisors in using **VigiFarmaco**, providing them with an initial automatic **MedDRA** encoding of the ADR descriptions in the online reports collected by **VigiFarmaco**, that the supervisors check and may correct or accept as it is. In this way, the encoding task, previously completely manual, becomes semi-automatic, reducing errors and the required time for accomplishing it. In spite of its first goal, **Magi-**

Coder has now evolved in an autonomous algorithm and software usable in all contexts where terms from a dictionary have to be recognized in a free narrative text. With respect to other solutions already available in literature and market, **MagiCoder** has been designed to be efficient and less computationally expensive, unsupervised, and with no need of training. **MagiCoder** uses stemming to be independent from singular/plural and masculine/feminine forms. Moreover, it uses string distance and other techniques to find best matching terms, discarding similar and non optimal terms.

With respect to the first version [8], we extended our proposal following several directions. First of all, we refined the procedure: **MagiCoder** has been equipped with some heuristic criteria and we started to address the problem of including auxiliary dictionaries (e.g., in order to deal with synonyms). **MagiCoder** computational complexity has been carefully studied and we will show that it is linear in the size of the dictionary (in this case, the number of LLTs in MedDRA) and the text description. We performed an accurate test of **MagiCoder** performances: by means of well-known statistical measures, we collected a significant set of quantitative information about the effective behavior of the procedure. We largely discuss some crucial key-points we met in the development of this version of **MagiCoder**, proposing short-time solutions we are addressing as work in progress, such as changes in stemming algorithm, considering synonyms, term filtering heuristics.

The paper is organized as follows. In Section 2 we provide some background notions and we discuss related work. In Section 3 we present the algorithm **MagiCoder**, by providing both a qualitative description and the pseudocode. In Section 4 we spend some words about the user interface of the related software tool. In Section 5 we explain the benchmark we developed to test **MagiCoder** performances and its results. Section 6 is devoted to some discussions. Finally, in Section 7 we summarize the main features of our work and sketch some future research lines.

2. Background and related work

2.1. Natural language processing and text mining in medicine

Automatic detection of adverse drug reactions from text has recently received an increasing interest in pharmacovigilance research. Narrative descriptions of ADRs come from heterogeneous sources: spontaneous reporting, Electronic Health Records, Clinical Reports, and social media. In [9–13] some NLP approaches have been proposed for the extraction of ADRs from text. In [14], the authors collect narrative discharge summaries from the Clinical Information System at New York Presbyterian Hospital. MedLEE, an NLP system, is applied to this collection, for identifying medication events and entities, which could be potential adverse drug events. Co-occurrence statistics with adjusted volume tests were used to detect associations between the two types of entities, to calculate the strengths of the associations, and to determine their cutoff thresholds. In [15], the authors report on the adaptation of a machine learning-based system for the identification and extraction of ADRs in case reports. The

role of NLP approaches in optimised machine learning algorithms is also explored in [16], where the authors address the problem of automatic detection of ADR assertive text segments from several sources, focusing on data posted by users on social media (Twitter and DailyStrenght, a health care oriented social media). Existing methodologies for NLP are discussed and an experimental comparison between NLP-based machine learning algorithms over data sets from different sources is proposed. Moreover, the authors address the issue of data imbalance for ADR description task. In [17] the authors propose to use association mining and Proportional Reporting Ratio (PRR, a well-know pharmacovigilance statistical index) to mine the associations between drugs and adverse reactions from the user contributed content in social media. In order to extract adverse reactions from on-line text (from health care communities), the authors apply the Consumer Health Vocabulary⁴ to generate ADR lexicon. ADR lexicon is a computerized collection of health expressions derived from actual consumer utterances, linked to professional concepts and reviewed and validated by professionals and consumers. Narrative text is preprocessed following standard NLP techniques (such as stop word removal, see Section 3.1). An experiment using ten drugs and five adverse drug reactions is proposed. The Food and Drug Administration alerts are used as the gold standard, to test the performance of the proposed techniques. The authors developed algorithms to identify ADRs from threads of drugs, and implemented association mining to calculate leverage and lift for each possible pair of drugs and adverse reactions in the dataset. At the same time, PRR is also calculated.

Other related papers about pharmacovigilance and machine learning or data mining are [18, 19]. In [20], a text extraction tool is implemented on the .NET platform for preprocessing text (removal of stop words, Porter stemming [21] and use of synonyms) and matching medical terms using permutations of words and spelling variations (Soundex, Levenshtein distance and Longest common subsequence distance [22]). Its performance has been evaluated on both manually extracted medical terms from summaries of product characteristics and unstructured adverse effect texts from Martindale (a medical reference for information about drugs and medicines) using the WHO-ART and MedDRA medical terminologies. A lot of linguistic features have been considered and a careful analysis of performances has been provided. In [23] the authors develop an algorithm in order to help coders in the subtle task of auto-assigning ICD-9 codes to clinical narrative descriptions. Similarly to MagiCoder, input descriptions are proposed as free text. The test experiment takes into account a reasoned data set of manually annotated radiology reports, chosen to cover all coding classes according to ICD-9 hierarchy and classification: the test obtains an accuracy of 77%.

⁴Available at <http://www.consumerhealthvocab.org>

MedDRA Level	MedDRA Term
SOC	Skin disorders
HLGT	Epidermal conditions
HLT	Dermatitis and Eczema
PT	Asteatotic Eczema
LLT	Itch

Table 1: MedDRA Hierarchy - an Example

2.2. MedDRA Dictionary

The Medical Dictionary for Regulatory Activities (**MedDRA**) [6] is a medical terminology used to classify adverse event information associated with the use of biopharmaceuticals and other medical products (e.g., medical devices and vaccines). Coding these data to a standard set of MedDRA terms allows health authorities and the biopharmaceutical industry to exchange and analyze data related to the safe use of medical products [24]. It has been developed by the International Conference on Harmonization (ICH); it belongs to the International Federation of Pharmaceutical Manufacturers and Associations (IFPMA); it is controlled and periodically revised by the **MedDRA** Maintenance And Service Organization (MSSO). **MedDRA** is available in eleven European languages and in Chinese and Japanese too. It is updated twice a year (in March and in September), following a collaboration-based approach: everyone can propose new reasonable updates or changes (due to effects of events as the onset of new pathologies) and a team of experts eventually decides about the publication of updates. **MedDRA** terms are organised into a hierarchy: the SOC (System Organ Class) level includes the most general terms; the LLT (Low Level Terms) level includes more specific terminologies. Between SOC and LLT there are three intermediate levels: HLGT (High Level Group Terms), HLT (High Level Terms), and PT (Preferred Terms).

The encoding of ADRs through **MedDRA** is extremely important for report analysis as for a prompt detection of problems related to drug-based treatments. Thanks to **MedDRA** it is possible to group similar/analogous cases described in different ways (e.g., by synonyms) or with different details/levels of abstraction.

Table 1 shows an example of the hierarchy: reaction *Itch* is described starting from *Skin disorders* (SOC), *Epidermal conditions* (HLGT), *Dermatitis and Eczema* (HLT), and *Asteatotic Eczema* (PT). Preferred Terms are Low Level Terms chosen to be representative of a group of terms. It should be stressed that the hierarchy is multi-axial: for example, a PT can be grouped into one or more HLT, but it belongs to only one primary SOC term.

3. MagiCoder: an NLP software for ADR automatic encoding

A natural language ADR description is a completely free text. The user has no limitations, she/he can potentially write everything: a number of online ADR

descriptions actually contain information not directly related to drug effects. Thus, an NLP software has to face and solve many issues: Trivial orthographical errors; Use of singular versus plural nouns; The so called “false positives”, i.e., syntactically retrieved inappropriate results, which are closely resembling to correct solutions; The structure of the sentence, i.e., the way an assertion is built up in a given language. Also the “intelligent” detection of linguistic connectives is a crucial issue. For example, the presence of a negation can potentially change the overall meaning of a description.

In general, a satisfactory automatic support of human reasoning and work is a subtle task: for example, the uncontrolled extension of the dictionary with auxiliary synonyms (see Section 6.2) or the naive ad hoc management of particular cases, can limit the efficiency and the desired of the algorithm. For these reasons, we carefully designed **MagiCoder**, even through a side-by-side collaboration between pharmacologists and computer scientists, in order to yield an efficient tool, capable to really support pharmacovigilance activities.

In literature, several NLP algorithms already exist, and several interesting approaches (such as the so called morpho-analysis of natural language) have been studied and proposed [7, 25, 26]. According to the described pharmacovigilance domain, we considered algorithms for the morpho-analysis and the part-of-speech (PoS) extraction techniques [7, 25] too powerful and general purpose for the solution of our problem. Indeed, in most cases ADR descriptions are written in a very succinct way, without using verbs, punctuation, or other lexical items, and introducing acronyms. Moreover, clinical and technical words are often not recognized correctly because not included in usual dictionaries. All these considerations limit the benefits of using morpho-analysis and PoS for our purposes.

Thus, we decided to design and develop an ad hoc algorithm for the problem we are facing, namely that of deriving **MedDRA** terms from narrative text and mapping segments of text in effective LLTs. This task has to be done in a very feasible time (we want that each interaction user/**MagiCoder** requires less than a second) and the solution offered to the expert has to be readable and useful. Therefore, we decided to ignore the *structure* of the narrative description and address the issue in a simpler way. Main features of **MagiCoder** can be summarized as follows:

- it requires *a single linear scan of the narrative description*: as a consequence, our solution is particularly efficient in terms of computational complexity;
- it has been designed and developed for the specific problem of mapping Italian text to **MedDRA** dictionary, but we claim the way **MagiCoder** has been developed is sound with respect to language and dictionary changes (see Section 7);
- the current version of **MagiCoder** is only based on the pure syntactical recognition of the text and it does not exploit any external synonym dictionary; in Section 4 we will discuss how synonyms may be used to in-

crease **MagiCoder** performances. In particular, we will discuss how a naïve approach to synonyms worsen computational and retrieval performances, while we will show through experimental results and empirical observations that a prudent and suitable use of an external dictionary produces an improvement of performances.

In this paper we consider the Italian context of Pharmacovigilance and, as a consequence, we will consider and process by **MagiCoder** textual descriptions written in Italian language. We will discuss the potentiality of **MagiCoder** on other languages and some preliminary results in Section 7.

3.1. *MagiCoder: overview*

The main idea of **MagiCoder** is that a single linear scan of the free-text is sufficient, in order to recognize **MedDRA** terms.

From an abstract point of view, we try to recognize, in the narrative description, *single words* belonging to LLTs, which do not necessarily occupy consecutive positions in the text. This way, we try to “reconstruct” **MedDRA** terms, taking into account the fact that in a description the reporter can permute or omit words. As we will show, **MagiCoder** has not to deal with computationally expensive tasks, such as taking into account subroutines for permutations and combinations of words (as, for example, in [20]).

We can distinguish five phases in the procedure that will be discussed in detail in Sections 3.1.1, 3.1.2, 3.1.3, 3.1.4, 3.1.5, respectively.

1. Definition of ad hoc data structures: the design of data structures is central to perform an efficient computation; our main data structures are hash tables, in order to guarantee an efficient access both to **MedDRA** terms and to words belonging to **MedDRA** terms.
2. Preprocessing of the original text: tokenization (i.e., segmentation of the text into syntactical units), stemming (i.e., reduction of words to a particular root form), elimination of computationally irrelevant words.
3. Word-by-word linear scan of the description and “voting task”: a word “votes” LLTs it belongs to. For each term voted by one or more words, we store some information about the retrieved syntactical matching.
4. Weights calculation: recognized terms are weighted depending on information about syntactical matching.
5. Sorting of voted terms and winning terms release: the set of voted term is pruned, terms are sorted and finally a solution (a set of winning terms) is released.

3.1.1. *Definition of ad hoc data structures*

The algorithm proceeds with a word-by-word comparison. We iterate on the preprocessed text and we test if a single word w , a token, occurs into one or many LLTs.

In order to efficiently test if a token belongs to one or more LLTs, we need to know which words belong to each term. The LLT level of MedDRA is actually a set of *phrases*, i.e., *sequences of words*. By scanning these sequences, we build a *meta-dictionary* of all the words which compose LLTs. As we will describe in Section 3.3, in $O(mk)$ time units (where m and k are the cardinality of the set of LLTs and the length of the longest LLT in MedDRA, respectively) we build a hash table having all the words occurring in MedDRA as keys, where the value associated to key w_i contains information about the set of LLTs containing w_i . This way, we can verify the presence in MedDRA of a word w encountered in the ADR description in constant time. We call this meta-dictionary DictByWord. We build a meta dictionary also from a stemmed version of MedDRA, to verify the presence of stemmed descriptions. We call it DictByWordStem. Finally, also the MedDRA dictionary is loaded into a hash table according to LLT identifiers and, in general, all our main data structures are hash tables.

We aim to stress that, to retain efficiency, we preferred exact string matching with respect to approximate string matching, when looking for a word into the meta dictionary. Approximate string matching would allow us to retrieve terms that would be lost in exact string matching (e.g., we could recognize misspelled words in the ADR description), but it would worsen the performances of the text recognition tool, since direct access to the dictionary would not be possible. We discuss the problem of retrieving syntactical variations of the same words and the problem of addressing orthographical errors in Section 7.

3.1.2. Preprocessing of the original ADR description

Given a natural language ADR description, the text has to be preprocessed in order to perform an efficient computation. We adopt a well-know technique such as tokenization [27]: a phrase is reduced to *tokens*, i.e., syntactical units which often, as in our case, correspond to words. A tokenized text can be easily manipulated as an enumerable object, e.g., an array. A *stop word* is a word that can be considered irrelevant for the text analysis (e.g., an article or an interjection). Words classified as stop-words are removed from the tokenized text. In particular, in this release of our software we decided to not take into account *connectives*, e.g., conjunctions, disjunctions, negations. The role of connectives, in particular of negation, is discussed in Section 6.

A fruitful preliminary work is the extraction of the corresponding *stemmed* version from the original tokenized and stop-word free text. Stemming is a linguistic technique that, given a word, reduces it to a particular kind of root form [21, 27]. It is useful in text analysis, in order to avoid problems such as missing word recognition due to singular/plural forms (e.g., hand/hands). In some cases, stemming procedures are able to recognize the same root both for the adjectival and the noun form of a word. Stemming is also potentially harmful, since it can generate so called “false positives” terms. A meaningful example can be found in Italian language. The plural of the word *mano* (in English, *hand*) is *mani* (in English, *hands*), and their stemmed root is *man*, which is also the stemmed version of *mania* (in English, *mania*). Several stemming

algorithms exist, and their impact on the performances of MagiCoder is discussed in Section 6.

3.1.3. Word-by-word linear scan of the description and voting task

MagiCoder scans the text word-by-word (remember that each word corresponds to a token) once and performs a “voting task”: at the i -th step, it marks (i.e., “votes”) with index i each LLT t containing the current (i -th) word of the ADR description. Moreover, it keeps track of the position where the i -th word occurs in t .

MagiCoder tries to find a word match both for the exact and the stemmed version of the meta dictionary and keeps track of the kind of match it has eventually found. It updates a flag, initially set to 0, if at least a stemmed matching is found in an LLT. If a word w has been exactly recognized in a term t , the match between the stemmed versions of w and t is not considered. At the end of the scan, the procedure has built a sub-dictionary containing only terms “voted” at least by one word. We call $\text{Voted}_{\text{LLT}}$ the sub-dictionary of voted terms.

Each voted term t is equipped with two auxiliary data structures, containing, respectively:

1. the positions of the *voting words* in the ADR description; we call voters_t this sequence of indexes;
2. the positions of the *voted words* in the MedDRA term t ; we call voted_t this sequence of indexes.

Moreover, we endow each voted term t with a third structure that will contain the *sorting criteria* we define below; we will call it weights_t .

Let us now introduce some notations we will use in the following. We denote as $t.\text{size}$ the function that, given an LLT t , returns the number of words contained in t (excluding the stop words). We denote as $\text{voters}_t.\text{length}$ (resp. $\text{voted}_t.\text{length}$) the function that returns the number of indexes belonging to voters_t (resp. voted_t). We denote as $\text{voters}_t.\text{min}$ and $\text{voters}_t.\text{max}$ the functions that return the maximum and the minimum indexes in voters_t , respectively.

From now on, sometimes we explicitly list the complete denomination of a terms: we will use the notation “*name*”(id), where “*name*” is the MedDRA description and *id* is its identifier, that is possibly used to refer to the term. Let us exemplify these notions by introducing an example. Consider the following ADR description: “anaphylactic shock (hypotension + cutaneous rash) 1 hour after taking the drug”. Words in it are numbered from 0 (anaphylactic) to 9 (drug). The complete set of data structures coming from the task is too big to be reported here, thus we focus only on two LLTs. At the end of the voting task, $\text{Voted}_{\text{LLT}}$ will include, among others, “Anaphylactic shock” (10002199) and “Anaphylactic reaction to drug” (10054844). We will have that $\text{voters}_{10002199} = [0, 1]$ (i.e., “anaphylactic” and “shock”) while $\text{voters}_{10054844} = [0, 9]$ (i.e., “anaphylactic” and “drug”). On the other hand,

$\text{voted}_{10002199} = [0, 1]$, revealing that both words in the term have been voted, while $\text{voted}_{10054844} = [0, 2]$, suggesting that only two out of three words in the term have been voted (in particular, “reaction” has not been voted). In this example all words in the description have been voted without using the stemming.

3.1.4. Weight calculation

After the voting task, selected terms have to be ordered. Notice that a purely syntactical recognition of words in LLTs potentially generates a large number of voted terms. For example, in the Italian version of MedDRA, the word “male” (in English, “pain”) occurs 3385 times.

So we have to: i) filter a subset of highly feasible solutions, by means of quantitative weights we assigns to candidate solutions; ii) choose a good final selection strategy in order to release a small set of final “winning” MedDRA terms (this latter point will be discussed in Section 3.1.5).

For this purpose, we define four criteria to assign “weights” to voted terms accordingly.

In the following, $\frac{1}{t.size}$ is a normalization factor (w.r.t. the length, in terms of words, of the LLT t). First three criteria have 0 as optimum value and 1 as worst value, while the fourth criterion has optimum value to 1 and it grows in worst cases.

Criterion one: Coverage

First, we consider how much part of the words of each voted LLT have not been recognized.

$$C_1(t) = \frac{t.size - \text{voted}_t.length}{t.size}$$

In the example we introduced before, we have that $C_1(10002199) = 0$ (i.e., all words of the terms have been recognized in the description) while $C_1(10054844) = 0.33$ (i.e., one word out of three has not been recognized in the description).

Criterion two: Type of Coverage

The algorithm considers whether a perfect matching has been performed using or not stemmed words. $C_2(\cdot)$ is simply a flag. $C_2(t)$ holds if stemming has been used at least once in the voting procedure of t , and it is valued 1, otherwise it is valued 0.

For example, $C_2(10002199) = 0$ and $C_2(10054844) = 0$.

Criterion three: Coverage Distance

The use of stemming allows one to find a number of (otherwise lost) matches. As side effect, we often obtain a quite large set of joint winner candidate terms. In this phase, we introduce a string distance comparison

between recognized words in the original text and voted LLTs. Among the possible string metrics, we use the so called pair distance [28], which is robust with respect to word permutation. Thus,

$$C_3(t) = pair(t, \bar{t})$$

where $pair(s, r)$ is the pair distance function (between strings s and r) and \bar{t} is the term “rebuilt” from the words in ADR description corresponding to indexes in $voters_t$.

For example, $C_3(10002199) = 0$ (i.e., the concatenation of the voters and the term are equal) and $C_3(10054844) = 12$.

Criterion four: Coverage Density

We want to estimate how an LLT has been covered.

$$C_4(t) = \frac{(voters_t.max - voters_t.min) + 1}{voted_t.length}$$

The intuitive meaning of the criterion is to quantify the “quality” of the coverage. If an LLT has been covered by nearby words, it will be considered a good candidate for the solution. This criterion has to be carefully implemented, taking into account possible duplicated voted words.

After computing (and storing) the weights related to the above criteria, for each voted term t we have the data structure $\mathbf{weights}_t = [C_1(t), C_2(t), C_3(t), C_4(t)]$, containing the weights corresponding to the four criteria. These weights will be used, after a first heuristic selection, to sort a subset of the syntactically retrieved terms.

Continuing the example introduced before, we have that $C_4(10002199) = 1$ while $C_4(10054844) = 5$. Thus, concluding, we obtain that $\mathbf{weights}_{10002199} = [0, 0, 0, 1]$ while $\mathbf{weights}_{10054844} = [0.33, 0, 12, 5]$.

3.1.5. Selection, ordering and release of winning terms

In order to provide an effective support to pharmacovigilance experts’ work, it is important to offer only a small set of good candidate solutions.

As previously said, the pure syntactical recognition of **MedDRA** terms into a free-text generates a possibly large set of results. Therefore, the releasing strategy has to be carefully designed in order to select onlt best suitable solutions. We will provide an heuristic selection, followed by a sorting of the survived voted terms; then we propose a release phase of solutions, further refined by a final heuristic criterium.

As a first step, we provide an initial pruning of the syntactically retrieved terms guided by the *ordered-phrases* heuristic criterium. In the ordered-phrases criterium we reintroduce the order of words in the narrative description as a selection discriminating factor. From the set of selected LLTs, we remove those

terms where voters (i.e., tokens in the original free text) appear in the ADR description in a relative order different from that of the corresponding voted tokens in the LLT. We do that *only* for those LLTs having voters that voted for more than one term.

Let us consider the following example. On the (Italian) narrative description “edema della glottide-lingua, parestesia al volto, dispnea” (in English, “edema glottis-tongue, facial paresthesia, dyspnoea”), the voting procedure of **Magi-Coder** finds, among the solutions, the **MedDRA** terms “Edema della glottide” (“Edema glottis”), “Edema della lingua” (“Edema tongue”), “Edema del volto” (“Edema face”), “Parestesia della lingua” (“Paresthesia tongue”), and “Dispnea” (“Dyspnoea”). The ordered-phrase criterium removes LLT “Parestesia della lingua” from the set of candidate solutions because “lingua” votes for two terms but in the narrative text it appears before than “parestesia” while in the LLT it appears after.

We call $\text{SelVoted}_{\text{LLT}}$ the set of voted terms after the selection by the ordered-phrase criterium. We proceed then by ordering $\text{SelVoted}_{\text{LLT}}$: we use a multiple-value sorting on elements in weights_t , for each $t \in \text{SelVoted}_{\text{LLT}}$. The obtained subdictionary is dubbed as $\text{SortedVoted}_{\text{LLT}}$ and it has possibly most suitable solutions on top.

After this phase, the selection of the “winning terms” takes place. The main idea is to select and return a subset of voted terms which “covers” the ADR description. We create the set $\text{Selected}_{\text{LLT}}$ as follows. We iterate on the ordered dictionary and for each $t \in \text{SortedVoted}_{\text{LLT}}$ we select t if all the following conditions hold:

1. t is completely covered, i.e., $C_1(t) = 0$;
2. t does not already belong to $\text{Selected}_{\text{LLT}}$;
3. t is not a prefix of another selected term $t' \in \text{SortedVoted}_{\text{LLT}}$;
4. t has been voted without stemming (i.e., $C_2(t) = 0$) or, for any $w_i \in \text{voters}_t$, w_i has not been covered (i.e., none term voted by w_i has been already selected) or w_i has not been exactly covered (i.e., only its stem has been recognized in some term t_1)⁵.

At this stage, we have a set of **MedDRA** terms which “covers” the narrative description. We further select a subset $\text{FinalVoted}_{\text{LLT}}$ of $\text{Selected}_{\text{LLT}}$ with a second heuristic, the *maximal-set-of-voters* criterium.

The maximal-set-of-voters criterium deletes from the solution those terms which can be considered “extensions” of other ones. For each pair of terms t_i

⁵In the implementation we add also the following thresholds: we choose only terms t such that $C_3(t) < 0.5$ and $C_4(t) < 3$. We extracted these thresholds by means of some empirical tests. We plan to eventually adjust them after some further performance tests.

and t_j , it checks if voters_{t_i} is a subset of voters_{t_j} (considered as sets of indexes). If it is the case, t_i is removed from $\text{Selected}_{\text{LLT}}$.

In **MagiCoder** we do not need to consider ad hoc subroutines to address permutations and combinations of words (as it is done, for example, in [20]). In Natural Language Processing, permutations and combinations of words are important, since in spoken language the order of words can change w.r.t. the formal structure of the sentences. Moreover, some words can be omitted, while the sentence still retains the same meaning. These aspects come for free from our voting procedure: after the scan, we retrieve the information that *a set of words covers a term* $t \in \text{Voted}_{\text{LLT}}$, but the order between words does not necessarily matter.

3.2. *MagiCoder: structure of the algorithm*

Figure 2 depicts the pseudocode of **MagiCoder**. We represent dictionaries either as sets of words or as sets of functions. We describe the main procedures and functions used in the pseudocode.

- Procedure *Preprocessing* takes the narrative description, performs tokenization and stop-word removal and puts it into an array of words.
- Procedures *CreateMetaDict* and *CreateMetaDictStem* get LLTs and create a dictionary of *words* and of their stemmed versions, respectively, which belong to LLTs, retaining the information about the set of terms containing each word.
- By the functional notation $\text{DictByWord}(w)$ (resp., $\text{DictByWordStem}(w)$), we refer to the set of LLTs containing the word w (resp., the stem of w).
- Function $\text{stem}(w)$ returns the stemmed version of word w .
- Function $\text{indx}_t(w)$ returns the position of word w in term t .
- stem_usage_t is a flag, initially set to 0, which holds 1 if at least a stemmed matching with the **MedDRA** term t is found.
- adr_clear , voters_t , voted_t are arrays and $\text{add}[A, l]$ appends l to array A , where l may be an element or a sequence of elements.
- C_i ($i = 1, 2, 3, 4$) are the weights related to the criteria defined in Section 3.1.4.
- Procedure $\text{sortby}(A, \{v_1, \dots, v_k\})$ performs the multi-value sorting of the array A based on the values of the properties v_1, \dots, v_k of its elements.
- Procedure $\text{prefix}(S, t)$, where S is a set of terms and t is a term, tests whether t (considered as a string) is *prefix* of a term in S . Dually, procedure $\text{remove_prefix}(S, t)$ tests if in S there are one or more prefixes of t , and eventually remove them from S .

- Function $mark(w)$ specifies whether a word w has been already covered (i.e., a term voted by w has been selected) in the (partial) solution during the term release: $mark(w)$ holds 1 if w has been covered (with or without stemming) and it holds 0 otherwise. We assume that before starting the final phase of building the solution (i.e., the returned set of LLTs), $mark(w) = 0$ for any word w belonging to the description.
- Procedures $ordered_phrases(S)$ and $maximal_voters(S)$, where S is a set of terms, implement *ordered-phrases* and *maximal-set-of-voters* criteria (defined in Section 3.1.5), respectively.
- Function $win(S, n)$, returns the first n elements of an ordered set S . If $|S| = m < n$, the function returns the complete list of ordered terms and $n - m$ nil values.

3.3. MagiCoder complexity analysis

Let us now conclude this section by sketching the analysis of the computational complexity of MagiCoder.

Let n be the input size (the length, in terms of words, of the narrative description). Let m be the cardinality of the dictionary (i.e., the number of terms). Moreover, let m' be the number of distinct words occurring in the dictionary and let k be the length of the longest term in the dictionary. For MedDRA, we have about 75K terms (m) and 17K unique words (m'). Notice that, reasonably, k is a small constant for any dictionary; in particular, for MedDRA we have $k = 22$. We assume that all update operations on auxiliary data structures require constant time $O(1)$.

Building meta-dictionaries DictByWord and DictByWordstems requires $O(km)$ time units. In fact, the simplest procedure to build these hash tables is to scan the LLT dictionary and, for each term t , to verify for each word w belonging to t whether w is a key in the hash table (this can be done in constant time). If w is a key, then we have to update the values associated to w , i.e., we add t to the set of terms containing w . Otherwise, we add the new key w and the associated term t to the hash table. We note that these meta-dictionaries are computed only once when the MedDRA dictionary changes (twice per year), then as many narrative texts as we want can be encoded without the need to rebuild them.

It can be easily verified that the voting procedure requires in the worst case $O(nm)$ steps: this is a totally conservative bound, since this worst case should imply that each word of the description appears in all the terms of the dictionary. A simple analysis of the occurrences of the words in MedDRA shows that this worst case never occurs: in fact, the maximal absolute frequency of a MedDRA word is 3937, and the average of the frequencies of the words is 19.1⁶. Thus, usually, real computational complexity is much less of this worst case.

⁶These values have been calculated excluding the stop-words and taking into account the stems of the words appearing in MedDRA.


```

Procedure MagiCoder( $D$  text, LLTDict dictionary,  $n$  integer)
Input:  $D$ : the narrative description;
        LLTDict: a data structure containing the MedDRA LLTs;
         $n$ : the maximum number of winning terms that have to be released by the procedure
Output: an ordered set of LLTs
DictByWord = CreateMetaDict(LLTDict);
DictByWordStem = CreateStemMetaDict(LLTDict);
 $adr\_clear$  = Preprocessing( $D$ );
 $adr\_length$  =  $adr\_clear.length$ ;
VotedLLT =  $\emptyset$ ;
/* for each non-stop-word in the description */
foreach ( $i \in [0, adr\_length - 1]$ ) do
    /* test whether the current word belongs to MedDRA */
    if  $adr\_clear[i] \in DictByWord$  then
        /* for each term containing the word */
        foreach  $t \in DictByWord(adr\_clear[i])$  do
            /* keep track of the index of the voting word */
            add[voters $t$ ,  $i$ ];
            /* keep track of the index of the recognized word in  $t$  */
            add[voted $t$ ,  $indx_t(adr\_clear[i])$ ];
            VotedLLT = VotedLLT  $\cup$   $t$ ;

    /* test if the current (stemmed) word belongs the stemmed MedDRA */
    if  $stem(adr\_clear[i]) \in DictByWordStem$  then
        foreach  $t \in DictByWordStem(stem(adr\_clear[i]))$  do
            /* test if the current term has not been exactly voted by the same word */
            if  $i \notin voters_t$  then
                add[voters $t$ ,  $i$ ];
                add[voted $t$ ,  $indx_t(adr\_clear[i])$ ];
                /* keep track that  $t$  has been covered by a stemmed word */
                 $stem\_usage_t = \text{true}$ ;
            VotedLLT = VotedLLT  $\cup$   $t$ ;

/* for each voted term, calculate the four weights of the corresponding criteria */
foreach  $t \in Voted_{LLT}$  do
    add[weights $t$ ,  $C_1(t), C_2(t), C_3(t), C_4(t)$ ]

/* filtering of the voted terms by the first heuristic criterium */
SelVotedLLT =  $orderd\_phrases(Voted_{LLT})$ ;
/* multiple value sorting of the voted terms */
SortedVotedLLT =  $sortBy(SelVoted_{LLT}, \{C_1, C_2, C_3, C_4\})$ ;
foreach  $t \in SortedVoted_{LLT}$  do
    foreach  $index \in voters_t$  do
        /* select a term  $t$  if it has been completely covered, its  $i$ -th voting word has not been covered
        or if its  $i$ -th voting word has been perfectly recognized in  $t$  and if  $t$  is not prefix of another
        already selected terms */
        if  $C_1(t) = 0$  AND ( $(stem\_usage_t = \text{false})$  OR ( $mark(adr\_clear(index)) = 0$ )) AND
         $t \notin Selected_{LLT}$  AND  $prefix(Selected_{LLT}, t) = \text{false}$  then
            mark( $adr\_clear(index)$ ) = 1;
            /* remove from the selected term set all terms which are prefix of  $t$  */
            SelectedLLT =  $remove\_prefix(Selected_{LLT}, t)$ ;
            SelectedLLT = SelectedLLT  $\cup$   $t$ ;

/* filtering of the finally selected terms by the second heuristic criterium */
FinalVotedLLT =  $maximal\_voters(Selected_{LLT})$ ;
winners =  $win(FinalVoted_{LLT}, n)$ ;
return winners

```

Figure 2: Pseudocode of MagiCoder

The computation of criteria-related weights requires $O(nm)$ time units. In particular: both criterion one and criterion two require $O(m)$ time steps; criterion three require $O(nm)$ (we assume to absorb the complexity of the pair

distance function); criterion four requires $O(nm)$ time units.

The subsequent multi-value sorting based on computed weights is a sorting algorithm which complexity can be approximated to $O(m \log m)$, based on the comparison of objects of four elements (i.e., the weights of the four criteria). Since the number of the criteria-related weights involved in the multi-sorting is constant, it can be neglected. Thus, the complexity of multi-value sorting can be considered to be $O(m \log m)$.

Finally, to derive the best solutions actually requires $O(nm)$ steps. The ordered-phrases criterium requires $O(nm)$; the maximal set of voters criterium takes $O(mn)$ time units.

Thus, we conclude that **MagiCoder** requires in the worst case $O(nm)$ computational steps. We again highlight that this is a (very) worst case scenario, while in average it performs quite better. Moreover, we did not take into account that each phase works on a subset of terms of the previous phase, and the size of these subset rapidly decreases in common application.

the selection phase works only on voted terms, thus, in common applications, on a subset of the original dictionary.

4. Software implementation: the user interface

MagiCoder has been implemented as a **VigiFarmaco** plug-in: people responsible for pharmacovigilance can consider the results of the auto-encoding of the narrative description and then revise and validate it. Figure 3 shows a screenshot of **VigiFarmaco** during this task. In the top part of the screen it is possible to observe the five sections composing a report. The screenshot actually shows the result of a human-**MagiCoder** interaction: by pressing the button “Autocodifica in MedDRA” (in English, “MedDRA auto-encoding”), the responsible for pharmacovigilance obtains a MedDRA encoding corresponding to the natural language ADR in the field “Descrizione” (in English, “Description”). Up to six solutions are proposed as the best MedDRA term candidates returned by **MagiCoder**: the responsible can refuse a term (through the trash icon), change one or more terms (by an option menu), or simply validate the automatic encoding and switch to the next section “Farmaci” (in English, “Drugs”). The maximum number of six terms to be shown has been chosen in order to supply pharmacovigilance experts with a set of terms extended enough to represent the described adverse drug reaction but not so large to be redundant or excessive.

We are testing **MagiCoder** performances in the daily pharmacovigilance activities. Preliminary qualitative results show that **MagiCoder** drastically reduces the amount of work required for the revision of a report, allowing the pharmacovigilance stakeholders to provide high quality data about suspected ADRs.

5. Testing **MagiCoder** performances

In this section we describe the experiments we performed to evaluate **MagiCoder** performances. The test exploits a large amount of manually revised reports we obtained from **VigiSegn** [4].

VigiFarmaco Segnalazione online ▼ Aiuto Profilo di Gabriele Pozzani Esci

Segnalazione online di sospetta reazione avversa da farmaci

Paziente

Reazione avversa

Farmaci

Dettagli aggiuntivi

Anteprima

Data di insorgenza 9 / Dicembre / 2015

Descrizione * gonfiore in sede di vaccinazione sx dal 5/11, febbre meno di 39.5 dal 21/11, vescicole, bolle presso la guancia dal 10/11

La descrizione può contenere fino a 255 caratteri

Autocodifica in MedDRA [Consulta il dizionario MedDRA](#)

Attenzione: verificare sempre la correttezza dei risultati dell'autocodifica. Le parole evidenziate in verde corrispondono a termini MedDRA LLT riconosciuti, selezionati e riportati tra l'elenco delle 6 reazioni codificate.

Guida alla compilazione

I campi contrassegnati con l'asterisco (*) sono obbligatori.

Per reazione avversa si intende un qualsiasi "effetto nocivo e non voluto conseguente all'uso di un medicinale".

Questo significa che vanno segnalate anche le reazioni avverse derivanti da errore terapeutico, abuso, misuso, uso off label, sovradosaggio ed esposizione professionale.

La descrizione della reazione avversa e dell'eventuale diagnosi devono avvenire nel modo più chiaro possibile.

MedDRA Reazione 1 * Vescicole ✖

MedDRA Reazione 2 Febbre ✖

MedDRA Reazione 3 Gonfiore in sede di vaccinazione ✖

MedDRA Reazione 4 Bolle ✖

MedDRA Reazione 5 nome della condizione ✖

MedDRA Reazione 6 nome della condizione ✖

Gravità * Grave Non grave

Criterio di esito Selezionare un criterio di esito

Figure 3: A partial screenshot of VigiFarmaco User Interface

We briefly recall two metrics we used to evaluate MagiCoder: *precision* and *recall*.

In statistical hypothesis and in particular in binary classification [29], two main kinds of errors are pointed out: *false positive errors* (FP) and *false negative errors* (FN). In our setting, these errors can be viewed as follows: a false positive error is the inopportune retrieval of a “wrong” LLT, i.e., a term which does not correctly encode the textual description; a false negative error is the failure in the recognition of a “good” LLT, i.e., a term which effectively encode (a part of) the narrative description and that would have been selected by a human expert. As dual notions of false positive and false negative, one can define *correct* results, i.e., *true positive* (TP) and *true negative* (TN): in our case, a true positive is a correctly returned LLT, and a true negative is an LLT which, correctly, has not been recognized as a solution.

Following the information retrieval tradition, the standard approach to sys-

Precision	$P = \frac{ \text{RelS} \cap \text{RetS} }{ \text{RetS} } = \frac{TP}{TP+FP}$
Recall	$R = \frac{ \text{RelS} \cap \text{RetS} }{ \text{RelS} } = \frac{TP}{TP+FN}$

Table 2: Performance and correctness measures

tem evaluation revolves around the notion of relevant and non-relevant solution (in information retrieval, a solution is represented by a document [29]). We provide here a straightforward definition of *relevant solution*. A relevant solution is a MedDRA term which correctly encode the narrative description provided to MagiCoder. A retrieved solution is trivially defined as an output term, independently from its relevance. We dub the sets of relevant solutions and retrieved solutions as RelS and RetS, respectively.

The evaluation of the false positive and the false negative rates, and in particular of the impact of relevant solutions among the whole set of retrieved solutions, are crucial measures in order to estimate the quality of the automatic encoding.

The *precision* (P), also called positive predictive value, is the percentage of retrieved solutions that are relevant. The *recall* (R), also called sensitivity, is the percentage of all relevant solutions returned by the system.

Table 2 summarizes formulas for precision and recall. We provide formulas both in terms of relevant/retrieved solutions and false positives, true positives and false negatives.

It is worth noting that the binary classification of solutions as relevant or non-relevant is referred to as the gold standard judgment of relevance. In our case, the gold standard has to be represented by a *human encoding* of a narrative description, i.e., a set of MedDRA terms choosen by a pharmacovigilance expert. Such a set is assumed to be definitively *correct* (only correct solutions are returned) and *complete* (all correct solutions have been returned).

5.1. Experiment about MagiCoder performances

To evaluate MagiCoder performances, we developed a benchmark, which automatically compares MagiCoder behavior with human encoding on already manually revised and validated ADR reports.

For this purpose, we exploited VigiSegn, a data warehouse and OLAP system that has been developed for the Italian Pharmacovigilance National Center [4]. This system is based on the open source business intelligence suite Pentaho⁷. VigiSegn offers a large number of *encoded* ADRs. The encoding has been manually performed and validated by experts working at pharmacovigilance centres. Encoding results have then been sent to the national regulatory authority, AIFA.

We performed a test composed by the following steps.

⁷<http://www.pentaho.com/>

Class	# chars	# reports	Common PT	FN	FP	R	P
1	0-20 chars	459	86%	8%	7%	86%	88%
2	21-40 chars	1012	68%	18%	14%	72%	75%
3	41-100 chars	1993	51%	25%	24%	61%	62%
4	101-255 chars	970	42%	24%	34%	58%	52%
5	>255 chars	11	33%	32%	35%	46%	45%

Table 3: First results of **MagiCoder** performances

1. We launch an ETL procedure through Pentaho Data Integration. Reports are transferred from **VigiSegn** to an ad hoc database **TestDB**. The dataset covers all the 4445 reports received, revised and validated during the year 2014 for the Italian region Veneto.
2. The ETL procedure extracts the narrative descriptions from reports stored in **TestDB**. For each description, the procedure calls **MagiCoder** from **VigiFarmaco**; the output, i.e., a list of **MedDRA** terms, is stored in a table of **TestDB**.
3. Manual and automatic encodings of each report are finally compared through an SQL query. In order to have two uniform data sets, we compared only those reports where **MagiCoder** recognized at most six terms, i.e., the maximum number of terms that human experts are allowed to select through the **VigiFarmaco** user interface. Moreover, we map each LLT term recognized by both the human experts and **MagiCoder** to its corresponding preferred term. Results are discussed below in Section 5.1.1.

5.1.1. Experiment: analysis of results

Table 3 shows the results of this first performance test. We group narrative descriptions by increasing length (in terms of characters). We note that reported results are computed considering terms at PT level. By moving to PT level, instead of using the LLT level, we group together terms that represent the same medical concept (i.e., the same adverse reaction). In this way, we do not consider an error when **MagiCoder** and the human expert use two different LLTs for representing the same adverse event. The use of the LLT level for reporting purpose and the PT level for analysis purpose is suggested also by **MedDRA** [6]. With *common PT* we mean the percentage of preferred terms retrieved by human reviewers that have been recognized also by **MagiCoder**. Reported performances are summarized also in Figure 4. Note that, false positive and false negative errors are required to be as small as possible, while common PT, recall, and precision have to be as large as possible.

MagiCoder behaves very well on very short descriptions (class 1) and on short ones (class 2). Recall and precision remain greater than 50% up to class 4. Notice that very long descriptions (class 5), on which performances drastically decrease, represent a negligible percentage of the whole set (less than 0.3%).

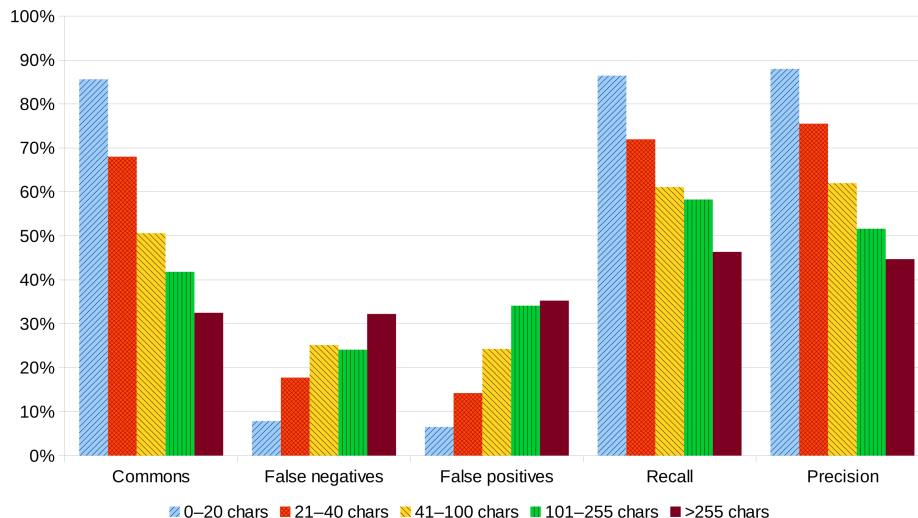


Figure 4: Graphical representation of **MagiCoder** performances

Some remarks are mandatory. It is worth noting that this test simply estimates how much, for each report, the **MagiCoder** behavior is similar to the manual work, without considering the effective quality of the manual encoding. Clearly, as a set of official reports, revised and sent to RNF, we assume to deal with an high-quality encoding: notwithstanding, some errors in the human encoding possibly occur. Moreover, the query we perform to compare manual and automatic encoding is, obviously, quantitative. For each **VigiSegn** report, the query is able to detect common retrieved terms and terms returned either by the human expert or by **MagiCoder**. It is not able to fairly test *redundancy* errors: human experts make some encoding choices in order to avoid repetitions. Thus, an LLT t returned by **MagiCoder** that has not been selected by the expert because redundant is not truly a false positive. As a significative counterpart, as previously said, we notice that some reports contain slightly human omissions/errors. This suggest the evidence that we are underestimating **MagiCoder** performances. See the next section for some simple but significative examples.

5.2. Examples

Table 4 provides some examples of the behavior of **MagiCoder**. We propose some free-text ADR descriptions from **TestDB** and we provide both the manual and the automatic encodings into LLT terms. We also provide the English translation of the natural language texts (we actually provide a quite straightforward *literal* translation).

#	Narrative Description	LLT Human Encoding	LLT MagiCoder Encoding
D1	Shock anafilattico (<i>ipotensione</i> + rash cutaneo) 1 h dopo assunzione x os del farmaco	<u>Shock anafilattico</u> ¹	Ipotensione , <u>Shock anafilattico</u> ¹
D2	gonfiore in sede di vaccinazione sx dal 5/11, febbre meno di 39,5 dal 21/11, vescicole, <i>bolle</i> presso la guancia dal 10/11	<u>Piressia</u> ² , <u>Vescicole</u> ³ , <u>Gonfiore in sede di vaccinazione</u> ¹	Bolle , <u>Febbre</u> ² , <u>Vescicole</u> ³ , <u>Gonfiore in sede di vaccinazione</u> ¹
D3	Reazione locale estesa, <i>dolore</i> locale; cefalea e febbre per due giorni	<u>Cefalea</u> ¹ , <u>Febbre</u> ² , <u>Reazione in sede di vaccinazione</u> ³	<u>Cefalea</u> ¹ , Dolore , <u>Febbre</u> ² , <u>Reazione locale</u> ³

Table 4: Examples of MagiCoder behavior

- D1: anaphylactic shock (hypotension + cutaneous rash) 1 hour after taking the drug.
- D2: swelling in vaccination location left from 11/5; temperature less than 39,5 from 11/21; vesicles, blisters around the cheek from 11/10.
- D3: extended local reaction, local pain, headache, fever for two days.

In Table 4 we use the following notations: $\underline{t_1}^n$ and $\underline{t_2}^n$ are two *identical* LLTs retrieved both by the human and the automatic encoding; $\overline{t_1}^n$ and $\overline{t_2}^n$ are two *semantically equivalent* or *similar* LLTs (i.e., LLTs with the same PT) retrieved by the human and the automatic encoding, respectively; we use bold type to denote terms that have been recognized by MagiCoder but that have not been encoded by the reviewer; we use italic type in D1, D2, D3 to denote text recognized only by MagiCoder. For example, in description D3, “cefalea” (in English, “headache”) is retrieved and encoded both by the human reviewer and MagiCoder; in description D2, ADR “febbre” (in English, “fever”) has been encoded with the term itself by the algorithm, whereas the reviewer encoded it with its synonym “piressia”; in D1, ADR “ipotensione” (in English, “hypotension”) has been retrieved only by MagiCoder.

To exemplify how the ordered phrase heuristic works, we can notice that in D2 MagiCoder did not retrieve the MedDRA term “Vescicole in sede di vaccinazione” (10069623), Italian for “Vaccination site vesicles”. It belongs to the set of the voted solutions (since $C_1(10069623) = 0$), but it has been pruned from the list of the winning terms by the ordered-phrase heuristic criterion.

6. Discussion

We discuss here some interesting points we met developing MagiCoder. We explain the choices we made and consider some open questions.

6.1. Stemming and performance of the NLP software

Stemming is a useful tool for natural language processing and text searching and classification. The extraction of the stemmed form of a word is a non-trivial operation, and algorithms for stemming are very efficient. In particular, stemming for Italian language is extremely critic: this is due to the complexity of language and the number of linguistic variations and exceptions.

For the first implementation of **MagiCoder** as **VigiFarmaco** plug-in, we used a robust implementation of the Italian stemming procedure⁸. The procedure takes into account subtle properties of the language; in addition of the simple recognition of words up to plurals and genres, it is able, in the majority of cases, to recognize an adjectival form of a noun by extracting the same syntactical root.

Despite the efficiency of this auxiliary algorithm, we noticed that the recognition of some **MedDRA** terms have been lost: in some sense, this stemming algorithm is too “aggressive” and, in some cases, counterintuitive. For example, the Italian adjective “psichiatrico” (in English, psychiatric) and its plural form “psichiatrici” have two different stems, “psichiatri” and “psichiatric”, respectively. Thus, in this case the stemmer fails in recognizing the singular and plural forms of the same word.

We then decided to adopt the stemming algorithm also used in Apache Lucene⁹, an open source text search engine library. This procedure is less refined w.r.t. the stemming algorithm cited above, and can be considered as a “light” stemmer: it simply elides the final vowels of a word¹⁰. This induces a conservative approach and a uniform processing of the whole set of **MedDRA** words. This is unsatisfactory for a general problem of text processing, but it is fruitful in our setting. We repeated the **MagiCoder** testing both with the classical and the light stemmer: in the latter case, we measure a global enhancement of **MagiCoder** performance. Regarding common retrieved preferred terms, we reveal an average enhancement of about 4%: percentages for classes 1, 2, 3, 4 and 5 move from 83%, 67%, 47%, 39%, 25%, respectively, to values in Table 3. It is reasonable to think that a simple stemming algorithm maintains the recognition of words up to plurals and genres, but in most cases, the recognition up to noun or adjectival form is potentially lost. Notwithstanding, we claim that it is possible to reduce this disadvantage thanks to the embedding in the dictionary of a reasonable set of synonyms of LLTs (see Section 6.2).

6.2. Synonyms

MagiCoder performs a pure syntactical recognition (up to stemming) of words in the narrative description: no semantical information is used in the current version of the algorithm. In written informal language, synonyms are frequently used. A natural evolution of our NLP software may be the addition of an Italian

⁸<http://snowball.tartarus.org/>

⁹<https://lucene.apache.org/>

¹⁰A refined stemmer acts in a more complex way, taking into account also the etymological source of the words.

thesaurus dictionary. This would appear a trivial extension: one could try to match MedDRA both with original words and their synonyms, and try to maximize the set of retrieved terms. We performed a preliminary test, and we observed a drastic deterioration of MagiCoder performances (both in terms of correctness and completeness): on average, common PT percentages decreases of 24%. The main reason is related to the nature of Italian language: synonymical groups include words related by figurative meaning. For example, among the synonyms of the word “faccia” (in English, “face”), one finds “viso” (in English “visage”), which is semantically related, but also “espressione” (in English, “expression”), which is not relevant in the considered medical context. Moreover, the use of synonyms of words in ADR text leads to an uncontrolled growth of the voted terms, that barely can be later dropped in the final terms release. Furthermore, the word-by-word recognition performed by MagiCoder, with the uncontrolled increase of the processed tokens (original words plus synonyms plus possible combinations), could induce a serious worsening of the computational complexity. Thus, we claim that this is not the most suitable way to address the problem and the designing of an efficient strategy to solve this problem is not trivial.

We are developing a different solution, working side-by-side with the pharmacovigilance experts. The idea, vaguely inspired by the Consumer Health Vocabulary (recalled in Section 2 and used in [17]), is to collect a set of *pseudo-LLTs*, in order to enlarge the MedDRA official terminology and to generate a new ADR lexicon. This will be done on the basis of frequently retrieved locutions which are semantically equivalent to LLTs. A pseudo LLT will be regularly voted and sorted by MagiCoder and, if selected, the software will release the official (semantically equivalent) MedDRA term. Notice that, conversely to the single word synonyms solution, each pseudo-LLT is related to one and only one official term: this clearly controls the complexity deterioration. Up to now, we added to the official MedDRA terminology a set of about 1300 locutions. We automatically generated such a lexicon by considering three nouns that frequently occur in MedDRA, “aumento”, “diminuzione” e “riduzione” (in English “increase”, “decrease”, and “reduction”, respectively) and their adjectival form. For each LLT containing one of these nouns (resp., adjectives) we generate an equivalent term taking into account the corresponding adjective (resp., noun).

This small set of synonyms induces a global improvement of MagiCoder performances on classes 4 and 5. For Class 4, both common retrieved PT percentage, precision and recall increase of 1%. For Class 5, we observe some significant increment: common retrieved PT moves from 33% to 37%; precision moves from 45% to 49%; recall moves from 46% to 55%.

Also false negative and false positive rates suggest that the building of the MedDRA-thesaurus is a promising extension. False negatives move from 23% to 22% for Class 4 and from 32% to 29% for Class 5. False positive percentage decrease of 1% both for Class 4 and Class 5.

Class 5, which enjoys a particular advantage from the introduction of the pseudo-LLTs, represents a small slice of the set of reports. Notwithstanding, these cases are very arduous to address, and we have, at least, a good evidence of the validity of our approach.

6.3. Connectives in the narrative descriptions

As previously said, in **MagiCoder** we do not take into account the *structure* of written sentences. In this sense, our procedure is radically different from those based on the so called part-of-speech (PoS) [30], powerful methodologies able to perform the morpho-syntactical analysis of texts, labeling each lexical item with its grammatical properties. PoS-based text analyzers are also able to detect and deal with logical connectives such as conjunctions, disjunctions and negations. Even if connectives generally play a central role in the logical foundation of natural languages, they have a minor relevance in the problem we are addressing: ADR reports are on average badly/hurriedly written, or they do not have a complex structure (we empirically noted this also for long descriptions). Notwithstanding, negation deserves a distinct consideration, since the presence of a negation can drastically change the meaning of a phrase. First, we evaluated the frequency of negation connectives in ADR reports: we considered the same sample exploited in Section 5.1, and we counted the occurrences of the words “non” (Italian for “not”) and “senza” (Italian for “without”)¹¹: we detected potential negations in 162 reports (i.e., only in the 3.5% of the total number, 4445). Even though negative sentences seem to be uncommon in ADR descriptions, the detection of negative forms is a short-term issue we plan to address. As a first step, we plan to recognize words that may represent negations and to signal them to the reviewer through the graphical UI. In this way, the software sends to the report reviewer an alert about the (possible) failure of the syntactical recognition.

6.4. On the selection of voted terms

As previously said, in order to provide an effective support to human revision work, it is necessary to provide only a small set of possible solutions. To this end, in the selection phase (described in Section 3.1.5), we performed drastic cuts on voted LLTs. For example, only completely covered LLTs can contribute to the set of winning terms. This is clearly a restrictive threshold, that makes completely sense in a context where at most six solutions can be returned. In a less restrictive setting, one can relax the threshold above and try to understand how to filter more “promising” solutions among partially covered terms. In this perspective, we developed a further criterion, the *Coverage Distribution*, based on assumptions we made about the structure of (Italian) sentences. The following formula simply sums the indexes of the covered words for $t \in \text{Voted}_{\text{LLT}}$:

$$C_5(t) = \sum_{i=0}^{\text{voted}_t.\text{length}-1} \text{voted}_t[i]$$

If $C_5(t)$ is small, it means that words in the first positions of term t have been covered. We defined $C_5(\cdot)$ to discriminate between possibly joint winning

¹¹The word “senza” does not necessarily imply a negation, thus we are clearly overestimating the presence of negations.

terms. Indeed, an Italian medical description of a pathology has frequently the following shape: *name of the pathology*+ “*location*” or *adjective*. Intuitively, we privilege terms for which the recognized words are probably the ones describing the pathology. The addition of $C_5(\cdot)$ (with the discard of condition $C_1(\cdot) = 0$ in the final selection) could improve the quality of the solution if a larger set of winning terms is admissible or in the case in which the complete ordered list of voted terms is returned.

7. Conclusions and future work

In this paper we proposed **MagiCoder**, a simple and efficient NLP software, able to provide a concrete support to the pharmacovigilance task, in the revision of ADR spontaneous reports. **MagiCoder** takes in input a narrative description of a suspected ADR and produces as outcome a list of **MedDRA** terms that “covers” the medical meaning of the free-text description. Differently from other BioNLP software proposed in literature, we developed an original text processing procedure. Preliminary results about **MagiCoder** performances are encouraging. Let us sketch here some ongoing and future work.

We are addressing the task to include ad hoc knowledges, as the **MedDRA**-thesaurus described in Section 6.2. We are also proving that **MagiCoder** is robust with respect to language (and dictionary) changes. The way the algorithm has been developed suggests that **MagiCoder** can be a valid tool also for narrative descriptions written in English. Indeed, the algorithm retrieves a set of words, which covers an LLT t , from a free-text description, only slightly considering the order between words or the structure of the sentence. This way, we avoid the problem of “specializing” **MagiCoder** for any given language. We plan to test **MagiCoder** on the English **MedDRA** and, moreover, we aim to test our procedure on different dictionaries (e.g., ICD-9 classification¹², WHO-ART¹³, SNOMED CT¹⁴). We are collecting several sources of manually annotated corpora, as potential testing platforms. Moreover, we plan to address the management of orthographical errors possibly contained in narrative ADR descriptions. We did not take into account this issue in the current version of **MagiCoder**. A solution could include an ad hoc (medical term-oriented) spell checker in **VigiFarmaco**, to point out to the user that she/he is doing some error in writing the current word in the free description field. This should drastically reduce users’ orthographical errors without heavy side effects in **MagiCoder** development and performances. Finally, we aim to apply **MagiCoder** (and its refinements) to different sources for ADR detection, such as drug information leaflets and social media [17, 31].

¹²<http://icd9cm.chrisendres.com/>

¹³<https://www.unc-products.com>

¹⁴<http://www.ihtsdo.org/snomed-ct/what-is-snomed-ct>

References

- [1] N. Arthur, A. Bentsi-Enchill, R. Couper, et al., The Importance of Pharmacovigilance - Safety Monitoring of Medicinal Products, World Health Organization, 2002.
- [2] J. Borg, G. Aislaitner, M. Pirozynski, S. Mifsud, Strengthening and rationalizing pharmacovigilance in the EU: where is Europe heading to? A review of the new EU legislation on pharmacovigilance, *Data Safety* 34 (3) (2011) 187–197.
- [3] L. Aagaard, J. Strandell, L. Melskens, P. Petersen, E. Holme Hansen, Global patterns of adverse drug reactions over a decade: analyses of spontaneous reports to vigibase, *Drug Safety* 35 (2012) 1171–1182.
- [4] A. Sabaini, Temporal data analysis and mining: A multidimensional approach and its application in a medical domain, Ph.D. thesis, Department of Computer Science, University of Verona - Italy (2015).
- [5] C. Combi, A. Masini, B. Oliboni, M. Zorzi, A logical framework for XML reference specification, *Lecture Notes in Computer Science (including sub-series Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9262 (2015) 258–267. doi:10.1007/978-3-319-22852-5_22.
- [6] International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH), MedDRA data retrieval and presentation: points to consider (2016).
- [7] D. Jurafsky, J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st Edition, Prentice Hall PTR, Upper Saddle River, NJ, USA, 2000.
- [8] M. Zorzi, C. Combi, R. Lora, M. Pagliarini, U. Moretti, Automagically encoding adverse drug reactions in MedDRA, in: 2015 International Conference on Healthcare Informatics, ICHI 2015, Dallas, TX, USA, October 21-23, 2015, pp. 90–99.
- [9] A. Bate, S. Evans, Quantitative signal detection using spontaneous ADR reporting, *Pharmacoepidemiology and Drug Safety* 18 (6) (2009) 427–436.
- [10] X. Wang, G. Hripcsak, M. Markatou, C. Friedman, Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: A feasibility study, *JAMIA* 16 (3) (2009) 328–337.
- [11] C. Friedman, Discovering novel adverse drug events using natural language processing and mining of the electronic health record, in: *Artificial Intelligence in Medicine*, Vol. 5651 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2009, pp. 1–5.

- [12] E. Aramaki, Y. Miura, M. Tonoike, T. Ohkuma, H. Masuichi, K. Waki, Extraction of adverse drug effects from clinical records, *Stud Health Technol Inform* 160 (Pt1) (2010) 739–743.
- [13] M. G. R. Reichley, P. Kilbridge, L. Noiro, R. N. R. W. Dunagan, T. Bailey, Natural language processing to identify adverse drug events, in: *AMIA Annu Symp Proc.*, 2008.
- [14] P. M. Kilbridge, L. A. Noiro, R. M. Reichley, T. C. Bailey, Computerized surveillance for adverse drug events in a pediatric hospital, *J Am Med Inform Assoc.* 16 (5) (09) 607–612.
- [15] H. Gurulingappa, A. Mateen-Rajput, L. Toldo, Extraction of potential adverse drug events from medical case reports, *Journal of Biomedical Semantics* 3 (15) (2012) 1–10.
- [16] A. Sarker, G. Gonzalez, Portable automatic text classification for adverse drug reaction detection via multi-corpus training, *Journal of Biomedical Informatics* 53 (2015) 196–207.
- [17] C. C. Yang, H. Yang, L. Jiang, M. Zhang, Social media mining for drug safety signal detection, in: *Proc. of the 2012 Int. Workshop on Smart Health and Wellbeing, SHB 2012*, 2012, pp. 33–40.
- [18] R. Harpaz, H. S. Chase, C. Friedman, Mining multi-item drug adverse effect associations in spontaneous reporting systems, *BMC Bioinformatics* 11 (S-9) (2010) S7.
- [19] N. Nissim, M. Boland, R. Moskovitch, N. Tatonetti, Y. Elovici, Y. Shahar, G. Hripcsak, An active learning framework for efficient condition severity classification, in: *Artificial Intelligence in Medicine (AIME'15)*, Vol. 9105 of *Lecture Notes in Computer Science*, Springer, 2015, pp. 13–24.
- [20] G. Dalhberg, Implementation and evaluation of a text extraction tool for adverse drug reaction information, Master's thesis, Uppsala University School of Engineering (2010).
- [21] M. Porter, An algorithm for suffix stripping, *Program* 14 (3) (1980) 130–137.
- [22] M. Collins, Tutorial: Machine learning methods in natural language processing, in: *Computational Learning Theory and Kernel Machines*, 16th Annual Conference on Computational Learning Theory, 2003, p. 655.
- [23] A. Coffman, N. Wharton, Clinical natural language processing: Auto-assigning ICD-9 codes, in: *Overview of the Computational Medicine Center's 2007 Medical Natural Language Processing Challenge*. Available online at <http://courses.ischool.berkeley.edu/i256/f09/FinalProjects>

- [24] P. Radhakrishna, Upversioning MedDRA dictionary - insights from a seasoned coder, *Data Basics* 20 (3) (2014) 1171–1182.
- [25] L. Bauer, *Introducing linguistic morphology*, Edinburgh University Press, Edinburgh, 2003.
- [26] K. Kishida, Technical issues of cross-language information retrieval: A review, *Inf. Process. Manage.* 41 (3) (2005) 433–455.
- [27] A. Clark, C. Fox, S. Lappin (Eds.), *The Handbook of Computational Linguistics and Natural Language Processing*, Blackwell Handbooks in Linguistics, John Wiley & Sons, 2010.
- [28] J. Piskorski, M. M. Sydow, String distance metrics for reference matching and search query correction, in: W. Abramowicz (Ed.), *Business Information Systems*, Vol. 4439 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2007, pp. 353–365.
- [29] C. D. Manning, P. Raghavan, H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA, 2008.
- [30] F. Dell’Orletta, Ensemble system for part-of-speech tagging, in: *Proceedings of EVALITA - Evaluation of NLP and Speech Tools for Italian*, Reggio Emilia, Italy, 2009, pp. 1–6.
- [31] A. Sarker, R. Ginn, A. Nikfarjam, K. O’Connor, K. Smith, S. Jayaraman, T. Upadhaya, G. Gonzalez, Utilizing social media data for pharmacovigilance: A review, *Journal of Biomedical Informatics* 54 (2015) 202 – 212.